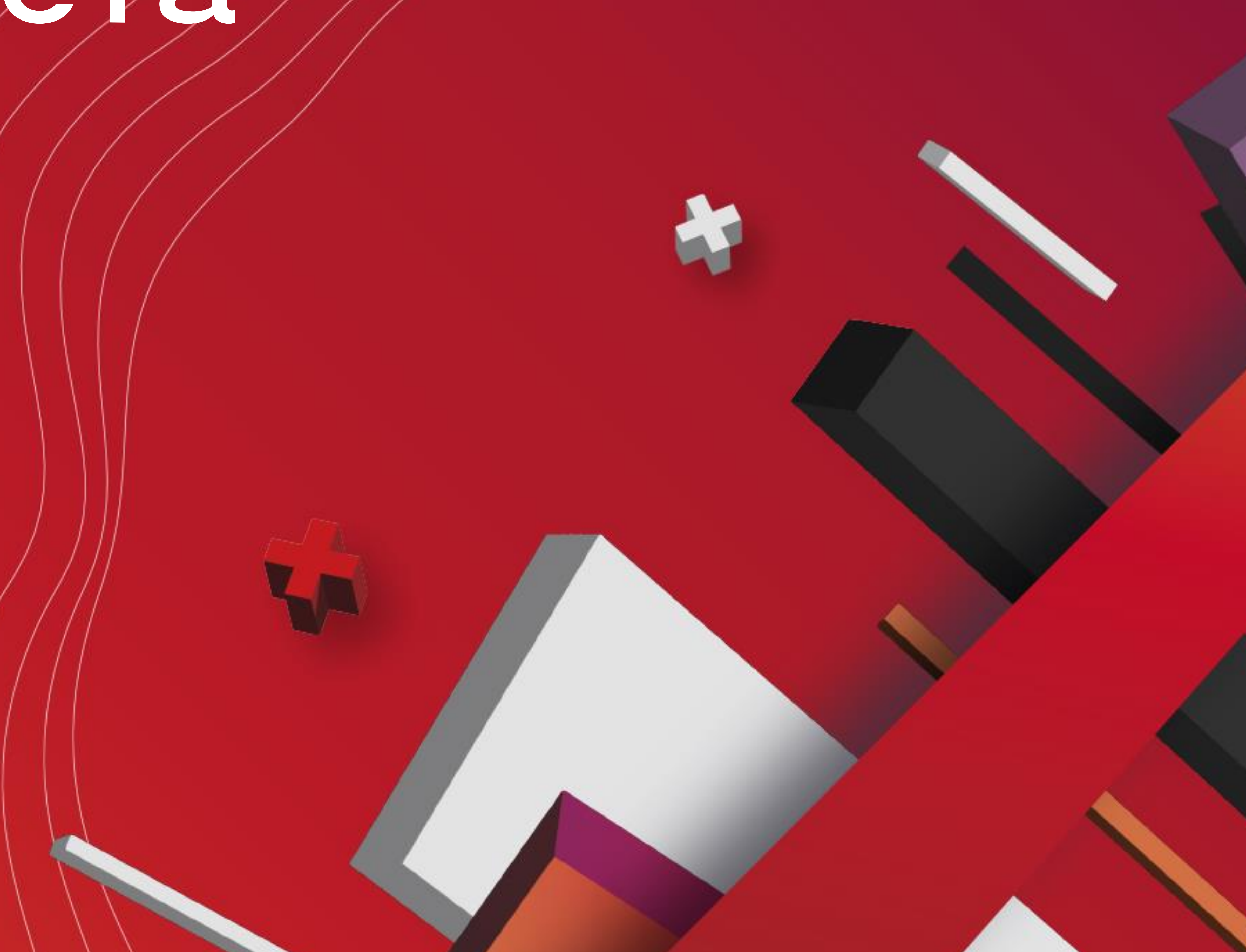
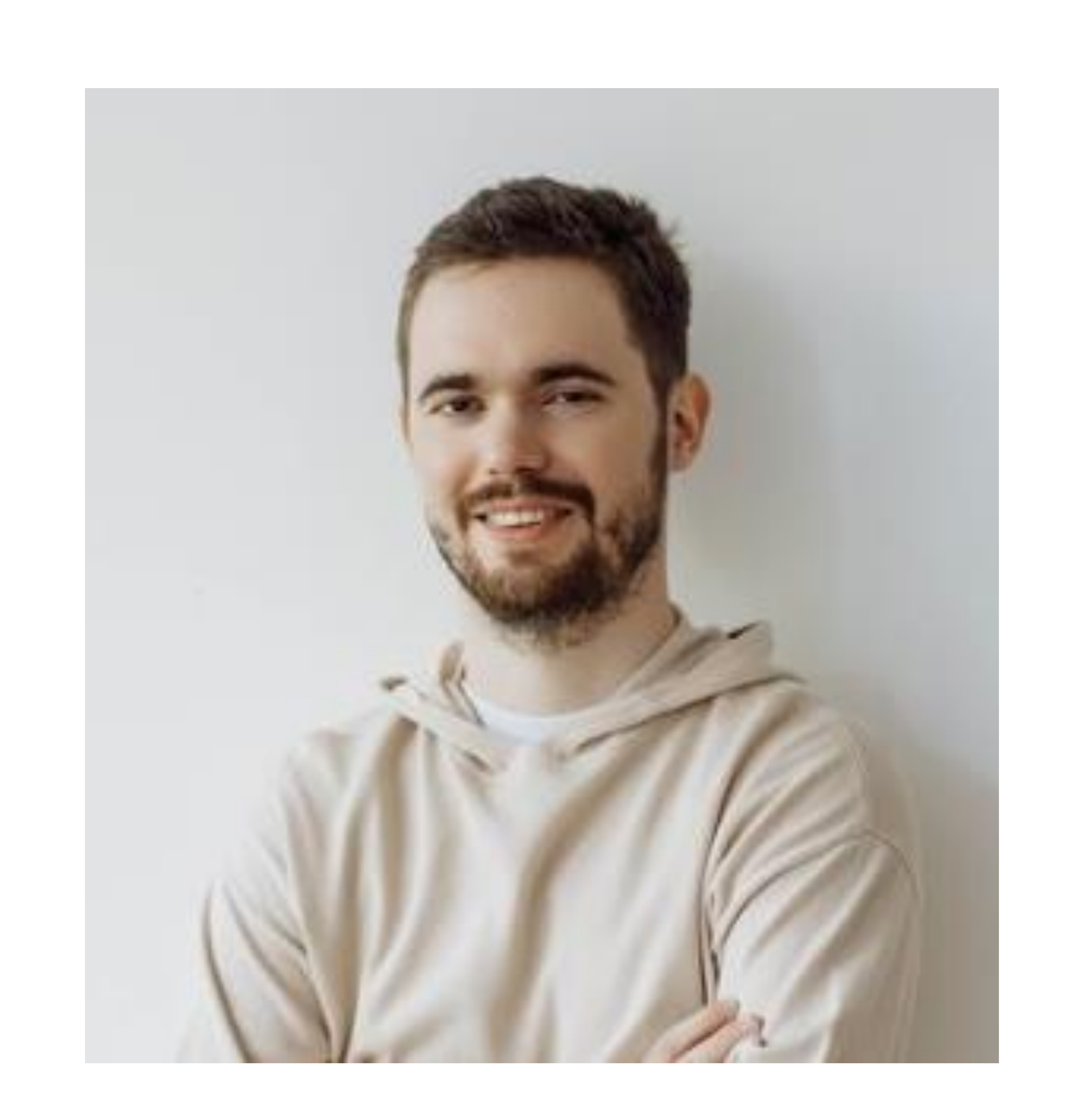


Кучумов Илья,
руководитель разработки
Товарного Поиска





Обо мне



- Руководитель отдела разработки Товарного Поиска в Яндексе
- В прошлом руководил качеством поиска в Лавке и разработкой Турбо-страниц



Яндекс

айфон 13

Найти

Поиск Картинки Видео Карты **Товары** Переводчик Все



Мобильные телефоны

По популярности | 1 Фильтры | Бренд ✕ | Цена | Линейка | Операционная система | Диагональ экрана

4,8
Смартфон Apple iPhone 13 128 ГБ, тёмная ночь
49 990 ₽
Лучшая цена
ещё 33 предложения

4,8
Смартфон Apple iPhone 13 256 ГБ RU, сияющая звезда
64 990 ₽
Лучшая цена
ещё 5 предложений

4,8
Смартфон Apple iPhone 13 128 ГБ RU, сияющая звезда
54 978 ₽
Лучшая цена
ещё 5 предложений

4,8
Смартфон Apple iPhone 13 512 ГБ RU, сияющая звезда
94 980 ₽
Лучшая цена
ещё 5 предложений

4,8
Смартфон Apple iPhone 13 512 ГБ RU, синий

4,8
Смартфон Apple iPhone 13 512 ГБ RU, розовый

4,8
Смартфон Apple iPhone 13 128 ГБ, синий

4,8
Смартфон Apple iPhone 13 256 ГБ, тёмная ночь



Смартфон Apple iPhone 13 128 ГБ, тёмная ночь

4,8 948 отзывов 94% рекомендуют

Цвет товара: тёмная ночь

Версия:

Ростест (EAC) | для других стран

Конфигурация памяти:

128 ГБ | 256 ГБ | 512 ГБ

Коротко о товаре

экран	6.1" (2532x1170) OLED 60 Гц
встроенная память	128 ГБ

Все характеристики

Предложение спонсора — Реклама

re-store.ru	75 990 ₽	В магазин

Цены в магазинах — Пожаловаться

ant-shop.ru ★ 4,6 732 отзыва	49 990 ₽ Лучшая цена	В магазин
berudevice.ru ★ 4,5 Нет отзывов	50 500 ₽	В магазин

4,8 948 отзывов 94% рекомендук

Цвет товара: тёмная ночь

Конфигурация памяти:

128 ГБ | 256 ГБ | 512 ГБ

Цены в магазинах

ant-shop.ru ★ 4,6 732 отзыва



Про что доклад

- Про построение базы цен Товарного Поиска
- Мы проверили, все работает и на других срезах
- Ключевая функциональность,
  высокие требования правильности
- Цены быстро меняются, невозможно кэшировать



План доклада

01 Общая архитектура базы Товарного Поиска

02 Подробно про парсинг и скачивание

03 Другие удачные применения парсинга

Что такое построение базы цен

20 млрд товаров

1,5 млн интернет-магазинов

~11% случайных товаров меняют цену каждые 24 часа

Ошибка цены — это плохо



Общая архитектура

Краулер — Поток html → Контент-система Робота — Поток данных → Хранилище

Контент-система Робота — html → Парсер

Парсер — data → Контент-система Робота

Хранилище — Полный индекс → Рантайм

Пользователь → Рантайм



Общая архитектура

Краулер — Поток html → Контент-система Робота —Поток данных→ Хранилище

Контент-система Робота — html → Парсер

Парсер — data → Контент-система Робота

Хранилище — Полный индекс → Рантайм

Пользователь → Рантайм



Общая архитектура

Краулер — Поток html → Контент-система Робота —Поток данных→ Хранилище

Контент-система Робота — html → Парсер

Парсер — data → Контент-система Робота

Хранилище — Полный индекс → Рантайм

Пользователь → Рантайм



Что такое парсинг

Чёрная Пятница | 0-0-24 Мегарассро... | -10% на комплект

Стиральная машина узкая Beko WSPE7H616W

beko

37 отзывов

Код товара: 2007 6822

Цвет: Белый

Максимальная загрузка — 7.5 кг
Мотор — инверторный
Максимальная скорость отжима — 1200 об/мин
Минипрограмма — 14 мин
Индикация времени до конца программы — Да
Расход воды за цикл — 55 л
Гарантия — 2 года
Страна — Россия

Все характеристики

Узкая стиральная машина Beko WSPE7H616W не занимает много места и подходит для установки в небольших помещениях. Есть 15 программ, среди которых ускоренные циклы и специализированные режимы для определённых типов вещей, например, для пуховиков и спортивных костюмов. СОВРЕМЕННЫЕ ТЕХНОЛОГИИ Сердце модели – инверторный мотор, который работает с низким уровнем...
Читать далее

Авторизуйтесь, чтобы снизить цену

28 999 ₽ ~~37 999~~

Посмотреть применённые скидки

Добавить в корзину

Нашли дешевле?

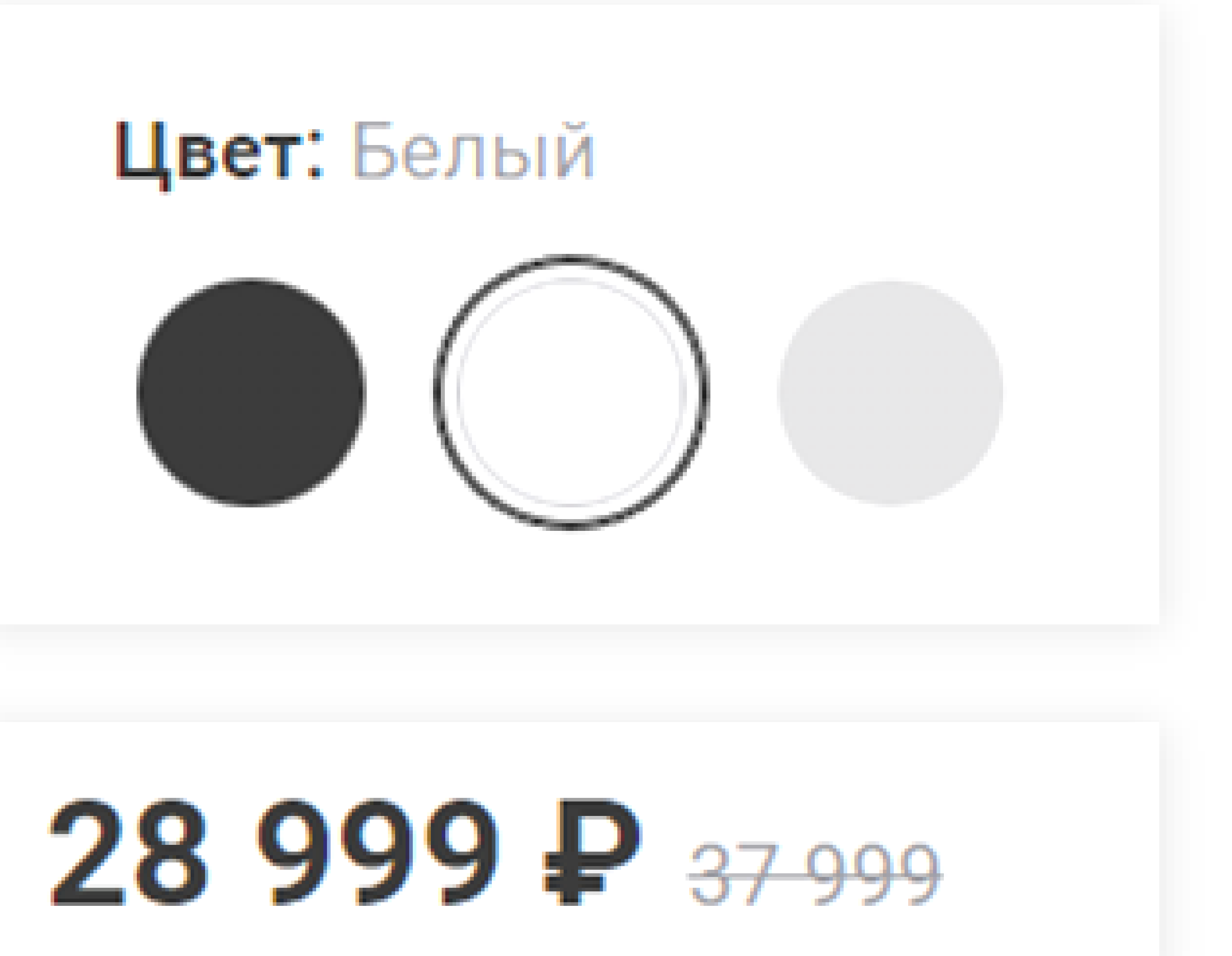
Кэшбэк +870 Бонусных рублей

Кредит от 2 832 ₽/мес

Рассрочка от 1 209 ₽/мес

Способы получения заказа

Самовывоз за 15 минут из 7 магазинов М.Видео Сегодня, бесплатно





Что такое парсинг

Электросамокат Ninebot KickScooter D18U, до 100 кг, черный

4.6 124 отзыва Характеристики 41 вопрос дизайн, качество сборки 23 370 человек интересовались за 2 месяца

Следить за снижением цены В избранное Сравнить

ЭКСКЛЮЗИВ

ЕЩЁ 17

Коротко о товаре

Максимальная нагрузка	100 кг
Запас хода	до 18 км без подзарядки
Максимальная скорость	25 км/ч
Мощность двигателя	250 Вт
Вес	14.9 кг
Особенности конструкции	амортизатор, складной
Материал рамы	алюминиевый сплав
Материал колес	резина
Аккумулятор	5100 мА·ч
Тип тормоза	ручной, комбинированный

Подробнее

Задать вопрос о товаре

Все товары Ninebot

СУПЕР ЦЕНЫ 19•21 октября 19 990 ₽

32 990 ₽

8 248 ₽ × 4 сплитом

660 баллов на Плюс

Самовывоз завтра, 19 октября — бесплатно

Курьером завтра, 19 октября — бесплатно

Доставка Яндекса со своего склада

Оплата онлайн

– 1 товар в корзине +

Маркет

2 мес 4 мес 6 мес

8 248 ₽ сегодня

и 24 742 ₽ потом

Оформить

18 окт	1 ноя	15 ноя	29 ноя
8 248 ₽	8 248 ₽	8 248 ₽	8 246 ₽

СУПЕР ЦЕНЫ 19•21 октября 19 990 ₽

32 990 ₽



Что такое парсинг

HTML →


```
{
  "title": "Стиральная машина",
  "price": 32999,
  "available": true,
  "main_image": "https:cdn.shop.ru/wash.jpg",
  "brand": "Самый лучший"
}
```




А можно ли просто?



Микроразметка

```
<meta itemprop="availability" content="https://schema.org/InStock" />
<meta itemprop="priceCurrency" content="RUB" />
<meta itemprop="price" content="82990" />
```

- \+ Быстро реализовать
- \+ Веб-мастера напрямую влияют
- \+ Покрытие — 50% товаров
- \+ Невозможно влиять со стороны парсера



Парсеры на селекторах

```
<div class="ProductCartFixedBlock__price">
    <span class="_current-price js--_current-price ">
        5990
    </span>
    <span class="ProductPrice__rouble __rouble">
        &#x20bd;
    </span>
</div>
```



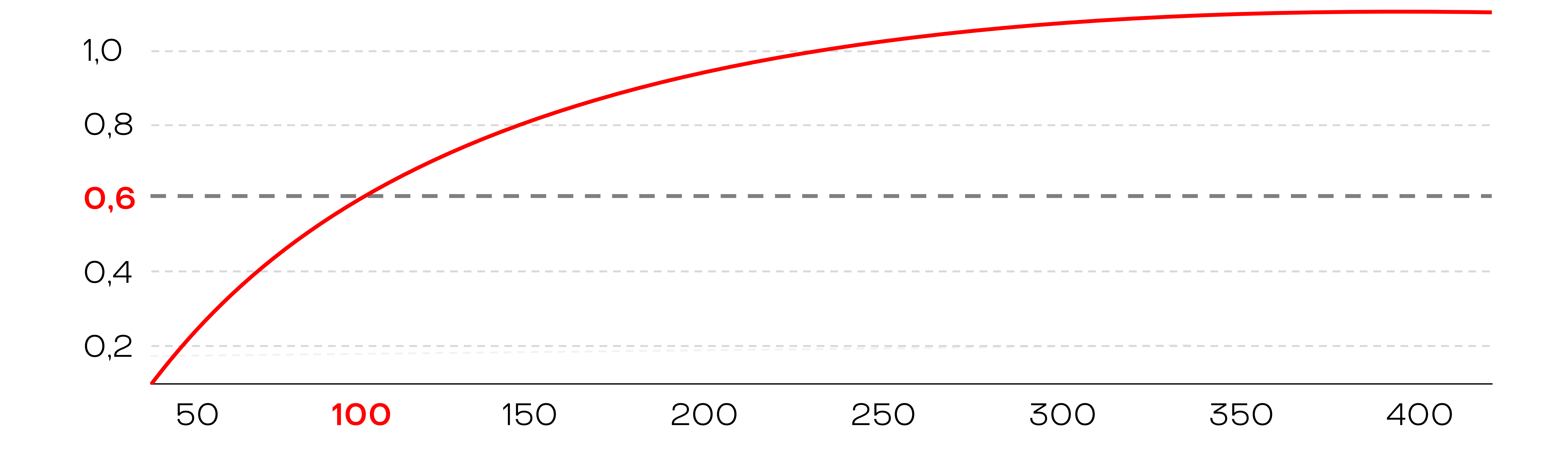

Парсеры на селекторах



Попытка накликать парсеры

Apple iPhone 13 Pro Max 256GB Dual SIM Silver купить...
best-magazin.com › Apple iPhone 13 Pro Max 256GB Dual SIM Silver (... ★ 4,9
Купить **Айфон 13 Про Макс 256** ГБ Dual SIM в Бест-магазин. ... Процессор **iPhone 13** Pro **Max** Dual SIM (**256**GB, Silver, 2 nano SIM) имеет 2 высокопроизводительных ядра (тактовая частота до 3,23 ГГц) плюс 4 ядра эффективности. Читать ещё
99 990 ₽







Попытка накликать парсеры

10 разработчиков

5 пиццц

+ Кола и чай
+ Большая переговорка

Мы написали ~~100~~ **80 парсеров за встречу**

~~80~~ **70 отправили в эксперимент**
(через месяц осталось 60)

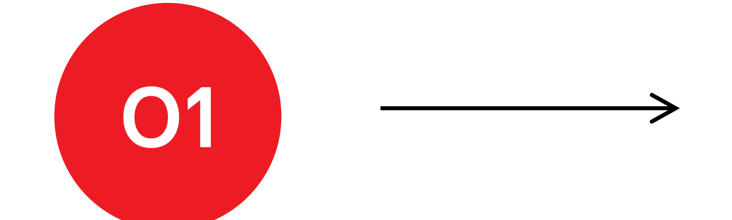


Парсеры на селекторах

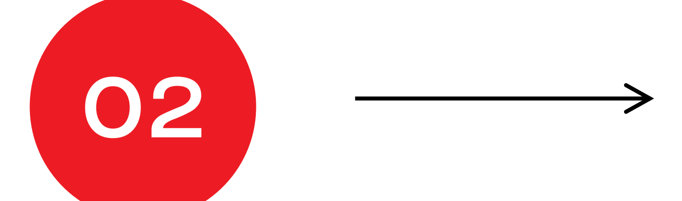
+ Возможность влиять на качество со стороны парсера
+ Быстро распарсить конкретный сайт
+ 1.5 млн. парсеров = около 130 рабочих человеко-лет накликивания
+ Нужна постоянная поддержка разработчиков



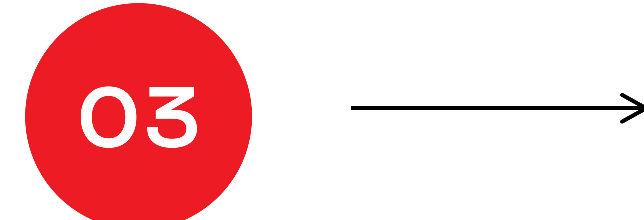
Парсеры краудом



Берем конкретный сайт



Толокер создает селекторы



Другие толокеры проверяют качество на семпле страниц

04

Если всё хорошо — добавляем в стейт парсера



Почему неидеальное решение?

+ Парсеры независимые

1.5 млн. парсеров = **1.1** млн. долларов

0,75 долларов/штука

30% случайных парсеров сломаются за месяц

- ❗ **Стоимость проверки 750 тыс. долларов (еженедельно)**
- ❗ **Расходы не делятся на другие срезы**

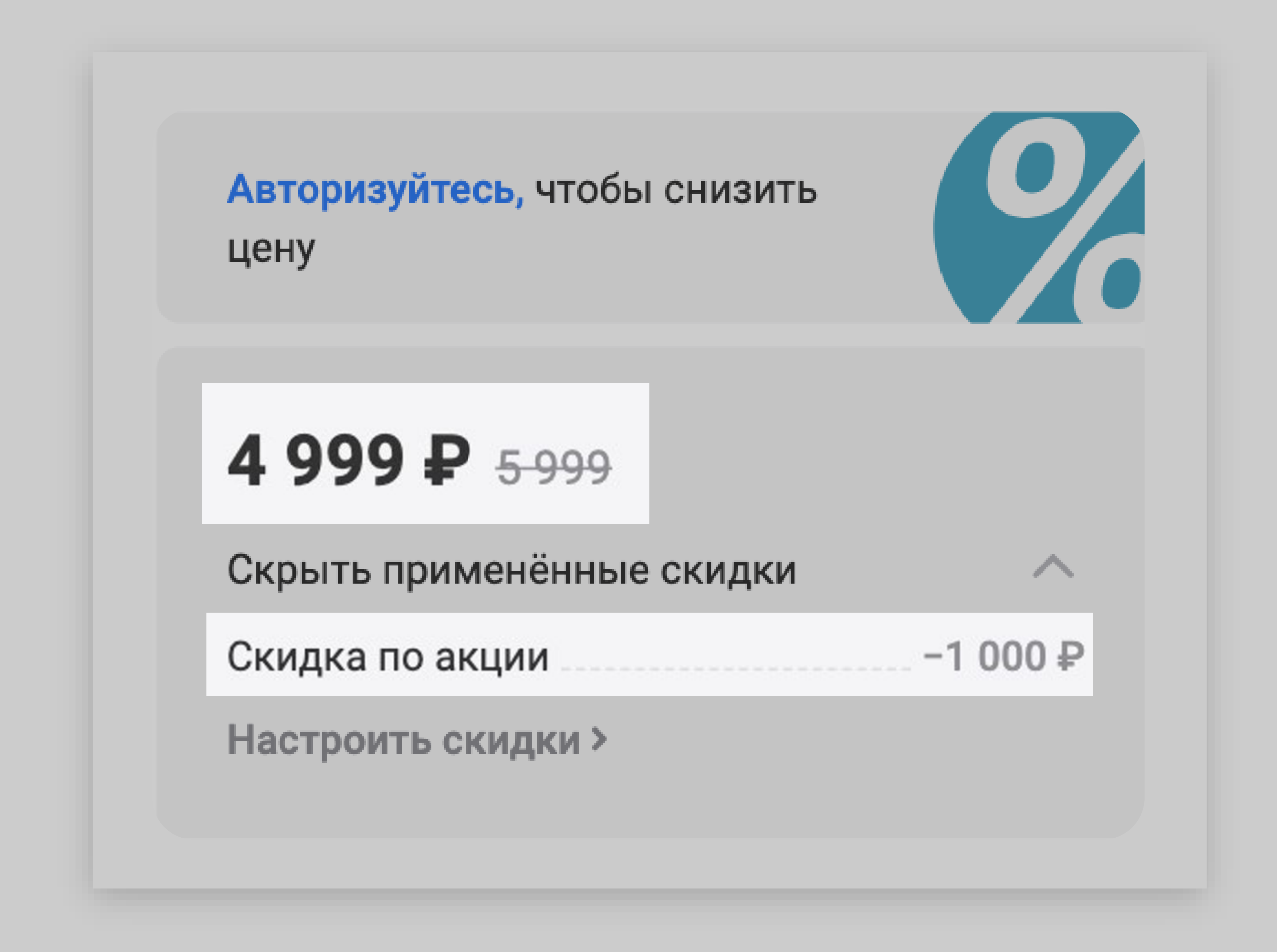


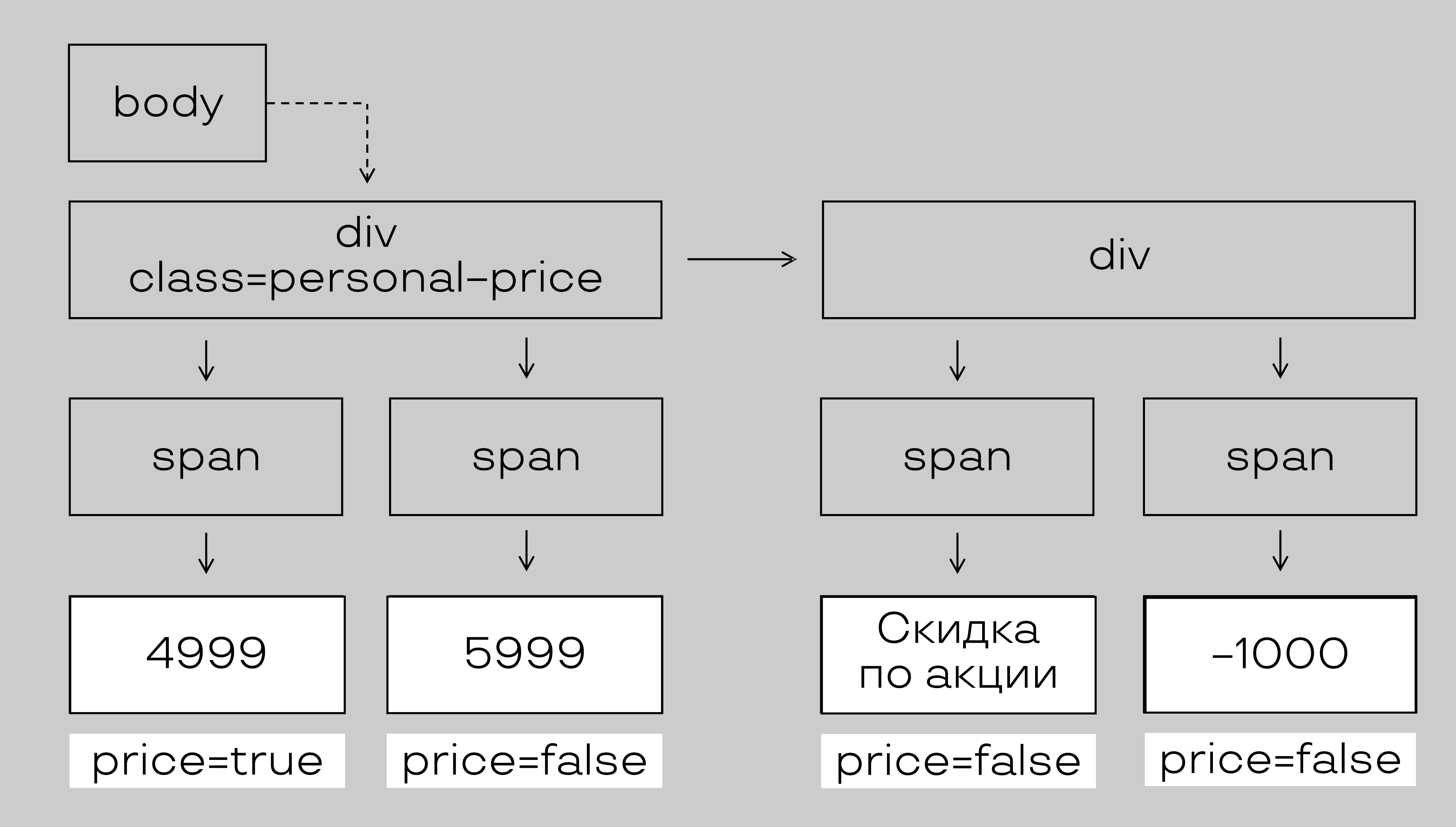
Саммари простых методов

- Простых рабочих эвристик по контенту нет
- Микроразметка — среднее покрытие, но быстро запустить
- Селекторы — хорошее покрытие, но дорого поддерживать



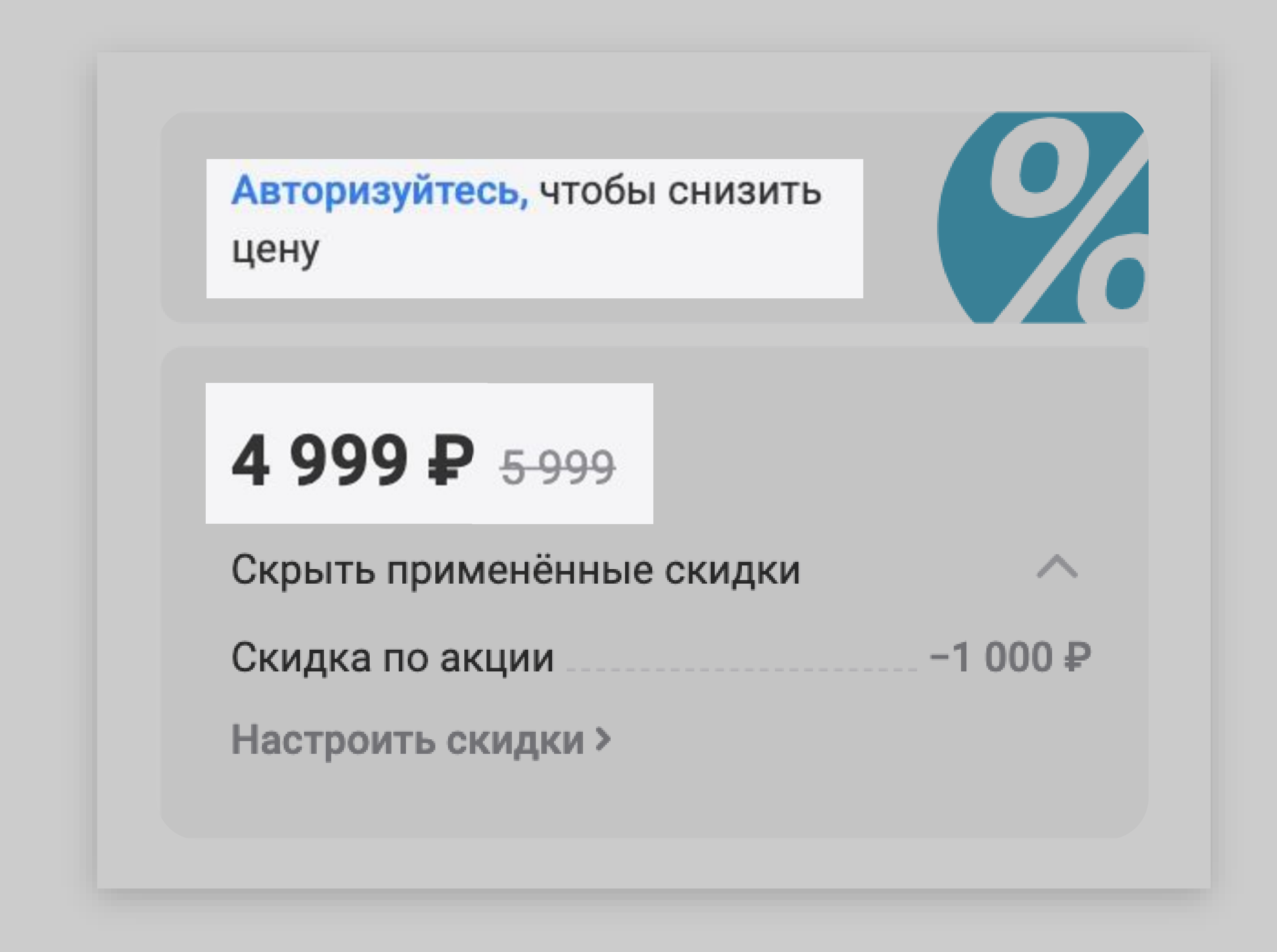
Классификация вершин DOM-дерева

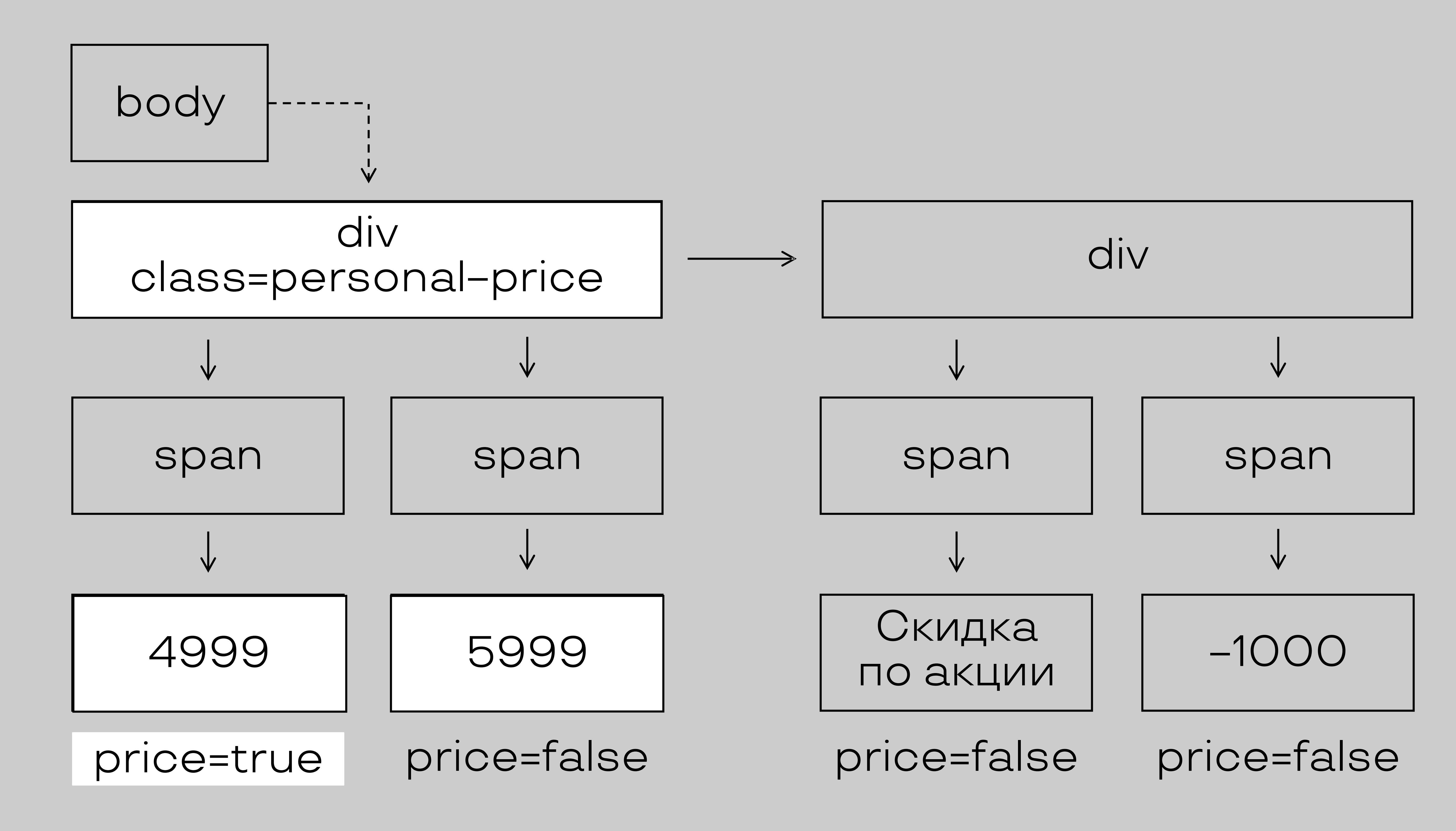






Текстовый фактор





Text = Авторизуйтесь, чтобы снизить цену | 4 999 || class=personal-price | 5999





Контекст вершины в дереве

- Тег родителя, дедушки, …
- Сколько параграфов (ссылок, картинок, …) у родителя, дедушки?
- А сколько еще таких же вершин в дереве?
- Количество вершин у родителя, дедушки, …



Позиция в дереве

body

div
class=params

div

div

span

span

span

span

span

span

цена

цвет

серый

размер

цена

27 200

span

12 900

classCountBefore=0
relPosWords=0,3

classCountBefore=1
relPosWords=0,8



Позиция в дереве

body

div
class=params

div

div

span
span
span
span
span
span
цена

цвет
серый
размер
цена
27 200
span
12 900

classCountBefore=0
relPosWords=0,3

classCountBefore=1
relPosWords=0,8



Необходимость оптимизаций

Краулер — Поток html → Контент-система Робота —Поток данных→ Хранилище

~100k RPS

Контент-система Робота — html → Парсер; Парсер — data → Контент-система Робота

Хранилище — Полный индекс → Рантайм

Пользователь → Рантайм

+ 1000 вершин
+ 800 факторов
+ ~10к CPU



Сужение потока до товарных страниц

Краулер — Поток html → Контент-система Робота — Поток данных → Хранилище

~100k RPS

Контент-система Робота — html ~10к RPS → Парсер — data → Контент-система Робота

Контент-система Робота — url, title → Классификатор товарности

Хранилище — Полный индекс → Рантайм ← Пользователь

+ F(url, title) = является ли страница товаром
+ Легкая DSSM'ка



Похостовое кэширование

Чёрная Пятница | 0-0-24 Мегарассро... | -10% на комплект

Стиральная машина узкая Beko WSPE7H616W

beko

37 отзывов

Код товара: 2007 6822

Цвет: Белый

Максимальная загрузка	7.5 кг
Мотор	инверторный
Максимальная скорость отжима	1200 об/мин
Минипрограмма	14 мин
Индикация времени до конца программы	Да
Расход воды за цикл	55 л
Гарантия	2 года
Страна	Россия

Все характеристики

Узкая стиральная машина Beko WSPE7H616W не занимает много места и подходит для установки в небольших помещениях. Есть 15 программ, среди которых ускоренные циклы и специализированные режимы для определённых типов вещей, например, для пуховиков и спортивных костюмов. СОВРЕМЕННЫЕ ТЕХНОЛОГИИ Сердце модели – инверторный мотор, который работает с низким уровнем...
Читать далее

Авторизуйтесь, чтобы снизить цену

28 999 ₽ ~~37 999~~
Посмотреть применённые скидки

Добавить в корзину

Нашли дешевле?

Кэшбэк **+870** Бонусных рублей

Кредит **от 2 832 ₽/мес**

Рассрочка **от 1 209 ₽/мес**

Способы получения заказа

Самовывоз за 15 минут из 7 магазинов М.Видео Сегодня, бесплатно

28 999 ₽ ~~37 999~~
Посмотреть применённые скидки



Похостовое кэширование

-2000 ₽ OfficeSuite
-20% на аксессуар
Windows 11 за 11 9...

Ноутбук Honor MagicBook X15 i5/16/512 Grey (BBR-WAH9)

18 отзывов

Код товара: 3006 2056

Диагональ/разрешение 15.6"/1920x1080 пикс.
Процессор Intel Core i5 10210U 1.6 ГГц
Оперативная память (RAM) 16 ГБ
Графический контроллер Intel UHD Graphics
Объем SSD 512 ГБ
Датчик отпечатков пальцев Да

Авторизуйтесь, чтобы снизить цену

Осталось мало! Сейчас этот товар смотрят 43 человека

69 999 ₽
Настроить скидки ›

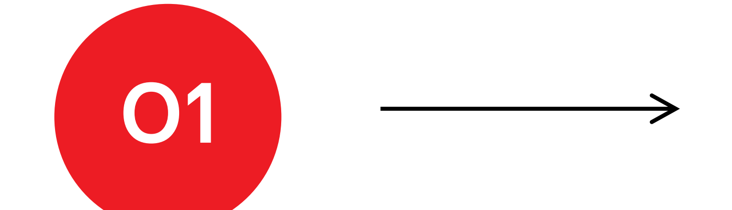
69 999 ₽
Настроить скидки ›



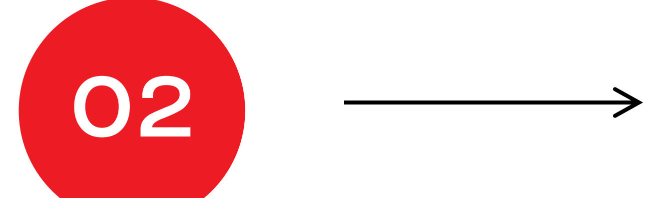
Давайте генерировать селекторы моделью



Алгоритм построения селекторов



Смотрим путь
от вершины
до корня



Выписываем
теги + класс
от вершины
до корня



Проверяем, что
все 99% вершин
с таким селектором
являются ценой

04

Сохраняем
селектор
для конкретного
сайта



Как используем

- Для каждого хоста (host, css_selector) → answer
- Файлик с миллионом записей
- Регулярно перестраиваем без участия крауда
- Экономия 60% железа



Как режется поток

100к RPS	**–90%** за счёт классификатора
10к RPS	**–50%** за счёт микроразметки
5к RPS	**–60%** за счёт кэширования
2к RPS	можно и распарсить



Почему актуальность цен важна

Электроника › Смартфоны и гаджеты ›

Мобильные телефоны

Сначала недорогие | 1 Фильтры | Бренд ✕ | Цена

Смартфон Apple iPhone XR 13 Pro 128GB Яблоко
40 000 ₽
Wildberries

Смартфон Apple iPhone 12 Pro max 128 ГБ, черный
48 897 ₽
OZON

Телефон Apple iPhone 13 128 ГБ Синий
51 990 ₽
applegod.ru

Смартфон Apple iPhone 12 ProMax 128 ГБ
52 197 ₽ ~~127 621 ₽~~
OZON



Давайте будем обходить базу раз в день (неделю?)



Поддержка актуальности

- Round robin не работает, прокачка всей базы — больше месяца
- Похостовые лимиты — серьёзная проблема
- 10M за день — это 116 RPS на конкретный сайт

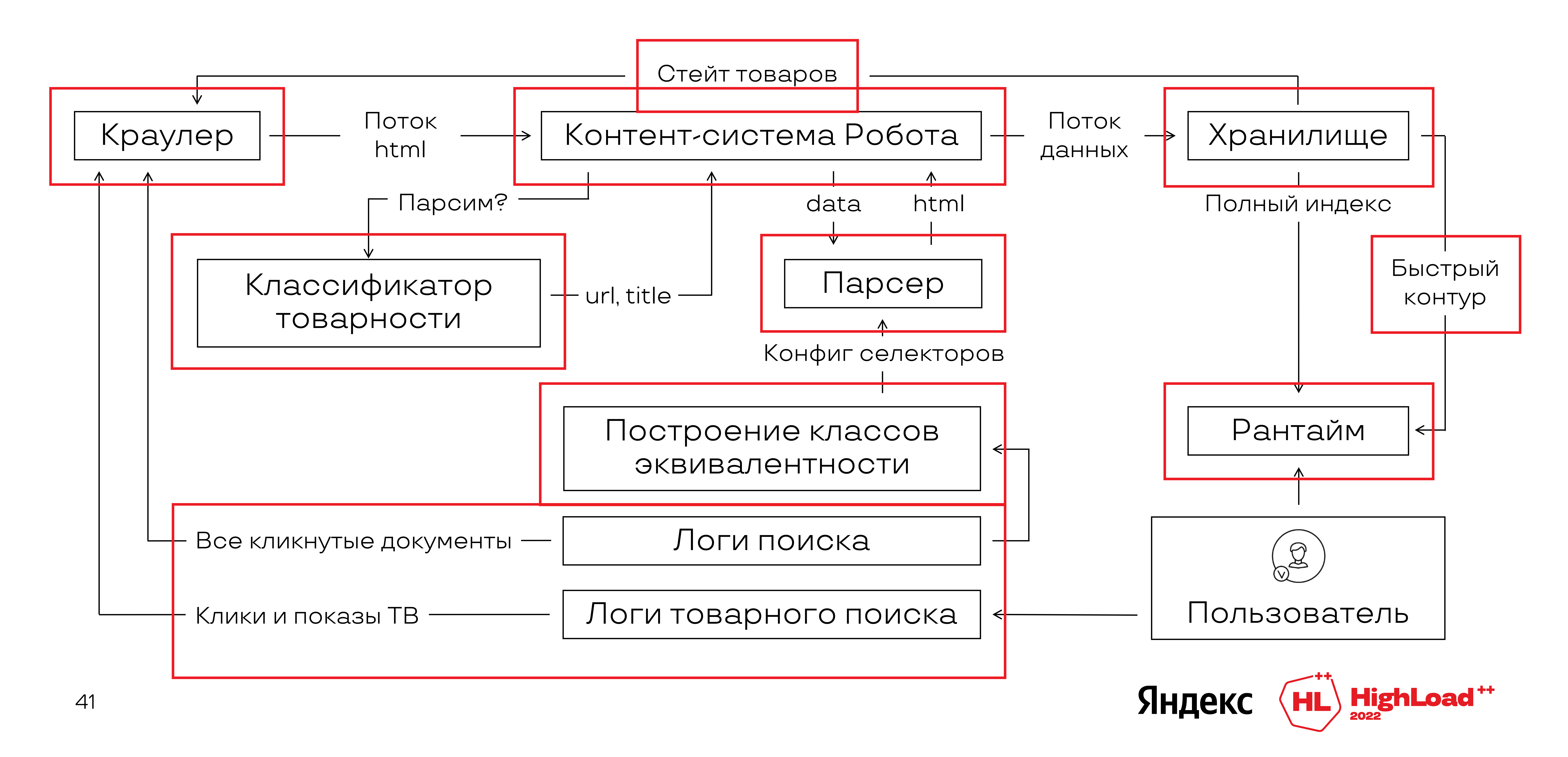


Эвристики скачивания

- Показы в поиске обходим всегда — покрываем популярные
- По (title, category, offer_age) предсказываем вероятность изменения цены с последнего скачивания
- Качаем самые вероятно изменившиеся
- Подбираем фильтр по возрасту, если нужно



Финальная схема



Что можно парсить ещё

Яндекс Практикум — профессия

Python-разработчик

Скидки на Госуслугах · Популярно

начало	с 1 ноября	цена за курс	138 600 ₽
длительность	9 месяцев	отзывы	112

в группе с наставником · для новичков · диплом · есть трудоустройство · есть бесплатная часть

Stepik — навык

"Поколение Python": курс для начинающих

Популярно

начало	в любой момент	цена за курс	бесплатно
длительность	3 часа	отзывы	28



Сложные характеристики товаров

Куртка мужская МОД 12032, темно-серый, 48/182

Цвет товара:

темно-серый | черный

Размер:

(48)182-96-80 | (50)182-100-84

Коротко о товаре

цвет товара	темно-серый
сезон	зима
состав	полиамид 100%
силуэт	прямой
застежка	молния, кнопки
детали	карманы, капюшон



Резюме

- Несколько подходов к парсингу структурированных данных
- Как поддерживать актуальность базы
- Архитектура базы Товарного Поиска
- Различные сферы применения алгоритмов парсинга



Резюме

- Несколько подходов к парсингу структурированных данных
- Как поддерживать актуальность базы
- Архитектура базы Товарного Поиска
- Различные сферы применения алгоритмов парсинга

Обратная связь и комментарии по докладу по ссылке



Кучумов Илья

руководитель разработки Товарного Поиска

@IlyaYndx

HL HighLoad++ 2022

Яндекс